

# Data clustering and noise undressing of correlation matrices

Matteo Marsili

*Istituto Nazionale per la Fisica della Materia (INFN), Trieste-SISSA Unit, V. Beirut 2-4, Trieste I-34014*  
(February 1, 2008)

We discuss a new approach to data clustering. We find that maximum likelihood leads naturally to an Hamiltonian of Potts variables which depends on the correlation matrix and whose low temperature behavior describes the correlation structure of the data. For random, uncorrelated data sets no correlation structure emerges. On the other hand for data sets with a built-in cluster structure, the method is able to detect and recover efficiently that structure. Finally we apply the method to financial time series, where the low temperature behavior reveals a non trivial clustering.

Statistical mechanics typically addresses the question of how structures and order arising from interactions in extended systems are dressed, and eventually destroyed, by stochastic – so-called thermal – fluctuations. The inverse problem, unraveling the structure of correlations from stochastic fluctuations in large data sets, has recently been addressed using ideas of statistical mechanics [1,2]. This is the case of data clustering problems, where the goal is to classify  $N$  objects, defined by  $D$  dimensional vectors  $\{\tilde{\xi}_i\}_{i=1}^N$ , in equivalence classes. Generally [1] the idea is 1) postulate a cost function, which depends on the data sample, for each structure and 2) consider the cost function as an Hamiltonian and study its thermal properties. Structures are identified by configurations  $\mathcal{S} = \{s_i\}_{i=1}^N$  of class indices, where  $s_i$  is the equivalence class to which object  $i$  belongs. Regarding  $s_i$  as Potts spins, a Potts Hamiltonian  $H_q = \sum_{i<j} J_{i,j} \delta_{s_i, s_j}$  has been recently proposed [2] as a cost function, with couplings  $J_{i,j}$  decreasing with the distance  $d_{i,j} = \|\tilde{\xi}_i - \tilde{\xi}_j\|$  between objects  $i$  and  $j$ . The underlying structure of data sets emerges as the clustering of Potts variables at low temperatures.

In this work we address the question of data clustering for time series. Rather than postulating the form of the Hamiltonian, we start from a statistical *ansatz* and invoke maximum likelihood and maximum entropy principles. In this way, the structure of the Hamiltonian arises naturally from the statistical *ansatz*, without the need of assumptions on its form. We study, by Monte-carlo method, this Hamiltonian for artificial time series: If time series are generated with some cluster structure  $\mathcal{S}^*$ , we find a low temperature phase which is dominated by cluster configurations close to  $\mathcal{S}^*$ . For random time series no low temperature phase is found. We also study time series of assets composing the S&P500 index, whose correlations have been the subject of much recent interest [3–5]. Correlation matrices of financial time series are of great practical interest. Indeed they are at the basis of risk minimization in the modern portfolio theory [6]. This states that, in order to reduce risk, the investment needs to be *diversified* (i.e. divided) on many uncorrelated assets. However the measure of correlation in finite samples was recently found to be affected by considerable

noise-dressing [3].

Our aim is to address the problem of revealing the structure of *bare* correlations hidden in a finite data set. Quite interestingly, our analysis of the S&P500 data set reveals a low temperature behavior dominated by few clusters of correlated assets with scale invariant properties. The thermal average over the relevant cluster structures provides a good fit of the financial correlations, which allows us to estimate the *noise-undressed* correlation matrix. Finally, we discuss several generalizations of our approach to generic data clustering.

The data  $\Xi = \{\tilde{\xi}_i\}_{i=1}^N$  is composed of  $N$  sets  $\tilde{\xi}_i = \{\xi_i(d)\}_{d=1}^D$  of  $D$  measures. These are normalized to zero mean  $\sum_d \xi_i(d)/D = 0$  and unit variance  $\sum_d \xi_i^2(d)/D = 1$ . We focus below on the case where  $\xi_i(d)$  is the normalized daily returns of asset  $i$  of the S&P500 index, in day  $d$  [7]. For the moment being, let us assume that  $\xi_i(d)$  are Gaussian variables. The reason is that we want to focus exclusively on pairwise correlations and the Gaussian model is the only one which is completely specified at this level. We shall discuss below how to apply the method when  $\xi_i(d)$  are not Gaussian. The key quantity of interest is the matrix

$$C_{i,j}(D) \equiv \frac{1}{D} \sum_{d=1}^D \xi_i(d) \xi_j(d). \quad (1)$$

The spectral properties of this matrix, for uncorrelated time series, are known exactly [8]. The spectrum of eigenvalues  $\lambda$  extends over an interval of size  $\sim N/D$  around  $\lambda = 1$ , as shown in Fig. 1a. The spectrum of eigenvalues of the S&P500 correlation matrix is also shown. The similarity of the two distributions for  $\lambda \approx 1$  suggests that significant noise-dressing due to finite  $D$  occurs [3]. The tail of the distribution ( $\lambda \gg 1$ ) implies that some correlation is however present. The structure of correlation was analyzed both by minimal spanning tree method [4] and by the method of ref. [2] in ref. [5].

In order to explain this correlation, Noh [9] proposed the *ansatz*

$$\xi_i(d) = \frac{\sqrt{g_{s_i}} \eta_{s_i}(d) + \epsilon_i(d)}{\sqrt{1 + g_{s_i}}}. \quad (2)$$

Here  $g_s > 0$  and  $s_i$  are integer variables (so-called Potts spins),  $\eta_s(d)$  and  $\epsilon_i(d)$  are *iid* gaussian variables with zero average and unit variance. In order to allow for totally uncorrelated sets, we allow  $s_i$  to take all integer values up to  $N$ . In Eq. (2) sets are correlated in clusters labeled by  $s$ . The  $s^{\text{th}}$  cluster is composed of  $n_s$  sets with *internal correlation*  $c_s$ , where

$$n_s = \sum_{i=1}^N \delta_{s_i, s}, \quad c_s = \sum_{i,j=1}^N C_{i,j} \delta_{s_i, s} \delta_{s_j, s}. \quad (3)$$

The correlation matrix generated by Eq. (2) for  $D \rightarrow \infty$  is  $C_{i,j} = (g_{s_i} \delta_{s_i, s_j} + \delta_{i,j}) / (1 + g_{s_i})$ . Its distribution of eigenvalues is simple: To each  $s$  with  $n_s \geq 1$  there correspond one eigenvalue  $\lambda_{s,0} = (1 + g_s n_s) / (1 + g_s)$  and  $n_s - 1$  eigenvalues  $\lambda_{s,1} = 1 / (1 + g_s)$ . Hence, large eigenvalues correspond to groups of many ( $n_s \gg 1$ ) sets. For  $D$  finite, we expect noise to lift degeneracies between  $\lambda_{s,1}$  but to leave the structure of large eigenvalues unchanged.

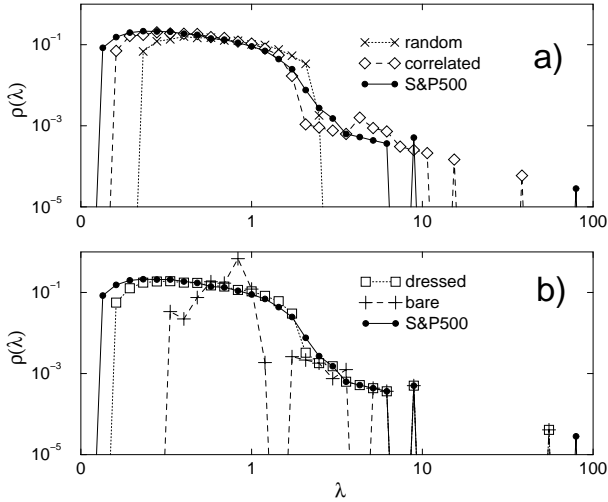


FIG. 1. **a)** Distribution of eigenvalues of the correlation matrices of S&P500 (full line •) random (dotted ×) and correlated (dashed ◇) time series. **b)** Comparison of the spectrum of the S&P500 correlation matrix (full line •) with noise-dressed (dotted □) and bare (dashed +) correlation matrices generated by Eq. (7).

In order to fit the data set  $\Xi$  with Eq. (2), let us compute the likelihood. This is the probability  $P(\Xi|\mathcal{S}, \mathcal{G})$  of observing the data  $\Xi$  as a realisation of Eq. (2) with structure  $\mathcal{S}$  and parameters  $\mathcal{G} = \{g_s\}_{s=1}^N$ , and it reads

$$P(\Xi|\mathcal{S}, \mathcal{G}) = \prod_{i=1}^N \prod_{d=1}^D \left\langle \delta \left( \xi_i(d) - \frac{\sqrt{g_{s_i}} \eta_{s_i}(d) + \epsilon_i(d)}{\sqrt{1 + g_{s_i}}} \right) \right\rangle$$

where the average is over all the  $\eta$ 's and  $\epsilon$ 's variables and  $\delta(x)$  is Dirac's delta function. Gaussian integration leads to  $P(\Xi|\mathcal{S}, \mathcal{G}) \propto e^{-DH\{\mathcal{S}, \mathcal{G}\}}$  with  $H\{\mathcal{S}, \mathcal{G}\} = \frac{1}{2} \sum_s [(1 + g_s)(n_s - \frac{g_s c_s}{1 + g_s n_s}) - n_s \ln(1 + g_s) + \ln(1 + g_s n_s)]$ . We fix the coupling strengths  $g_s$  by likelihood maximization  $\frac{\partial H}{\partial g_s} = 0$  for all  $s$ , which yields

$$\hat{g}_s = \frac{c_s - n_s}{n_s^2 - c_s} \quad (4)$$

for  $n_s > 1$  and  $\hat{g}_s = 0$  for  $n_s \leq 1$ . Note that for uncorrelated sets  $C_{i,j} = \delta_{i,j}$  we have  $c_s = n_s \forall s$  and hence  $\hat{g}_s = 0$ . The coupling strength  $\hat{g}_s$  instead diverges for totally correlated sets ( $C_{i,j} = 1$ ) because  $c_s = n_s^2$ . Using Eq. (4) we find that the likelihood of structure  $\mathcal{S}$  under ansatz (2) takes the form  $P(\Xi|\mathcal{S}) \propto e^{-DH_c}$ , where

$$H_c\{\mathcal{S}\} = \frac{1}{2} \sum_{s: n_s > 0} \left[ \log \frac{c_s}{n_s} + (n_s - 1) \log \frac{n_s^2 - c_s}{n_s^2 - n_s} \right] \quad (5)$$

The ground state  $\mathcal{S}_0$  of  $H_c$  yields the maximum likelihood fit with Eq. (2). This would probably take the ansatz (2) too seriously. In general, it is preferable to consider probabilistic solutions  $P\{\mathcal{S}\}$  and, following ref. [1], we invoke the maximum entropy principle: Among all distributions  $P\{\mathcal{S}\}$  with the same average log-likelihood, we select that which has maximal entropy. This, as usual, leads to the Gibbs distribution  $P\{\mathcal{S}\} \propto e^{-\beta H_c\{\mathcal{S}\}}$  where the inverse temperature  $\beta$  arises as a Lagrange multiplier.

The Hamiltonian  $H_c$  depends implicitly on the Potts spins  $s_i$  through the cluster variables  $n_s$  and  $c_s$  of Eq. (3). Unlike the Potts Hamiltonian  $H_q$ , the dependence on  $\delta_{s_i, s_j}$  is non-linear and it is modulated by  $C_{i,j}$ . For  $s_i \neq s_j$  for all  $i \neq j$  we have  $n_s = c_s = 1$  for all  $s$  and hence  $H_c = 0$ . This state is representative of the high temperature ( $\beta \rightarrow 0$ ) limit. The low temperature physics of  $H_c$  is instead non-trivially related to the correlation matrix  $C_{i,j}$ . Note, that the ferromagnetic state  $s_i = 1, \forall i$ , which dominates as  $\beta \rightarrow \infty$  in clustering methods based on Potts models [2], is in general not the ground state of  $H_c$ . Intuitively we expect that, if the model of Eq. (2) is reasonable,  $H_c$  should have a well defined ground state and low temperature phase which is energetically dominated by this state. In these cases, as in ref. [2], we expect a thermal phase transition [10].

In order to study the properties of  $H_c$  we resort to Montecarlo (MC) method by Metropolis algorithm [11]. This, at equilibration, allows us to sample the Gibbs distribution  $P\{\mathcal{S}\}$  and compute average quantities, such as the internal energy  $E_\beta = \langle H_c \rangle_\beta$  where  $\langle \dots \rangle_\beta$  stands for thermal average. In order to detect the occurrence of spontaneous magnetization – which occurs if the  $s_i$  lock into energetically favourable configurations at low temperature – we measure the autocorrelation function

$$\chi(t, \tau) = \frac{\sum_{i < j} \delta_{s_i(t), s_j(t)} \delta_{s_i(t+\tau), s_j(t+\tau)}}{\sum_{i < j} \delta_{s_i(t), s_j(t)}}. \quad (6)$$

This quantity tells us how many pairs of sites belonging to the same cluster at time  $t$  are still found in the same cluster after  $\tau$  MC steps. For  $t$  large enough,  $\chi$  becomes a function of  $\tau$  only. This function decreases rapidly to a plateau value  $\chi_\beta = \langle \chi(t, \tau) \rangle_\beta$  for  $t \gg \tau \gg 1$ . Clearly

$\chi_\beta \simeq 0$  implies that no persistent structure is present whereas, at the other extreme,  $\chi_\beta = 1$  implies that all sites are locked in a persistent structure of clusters.

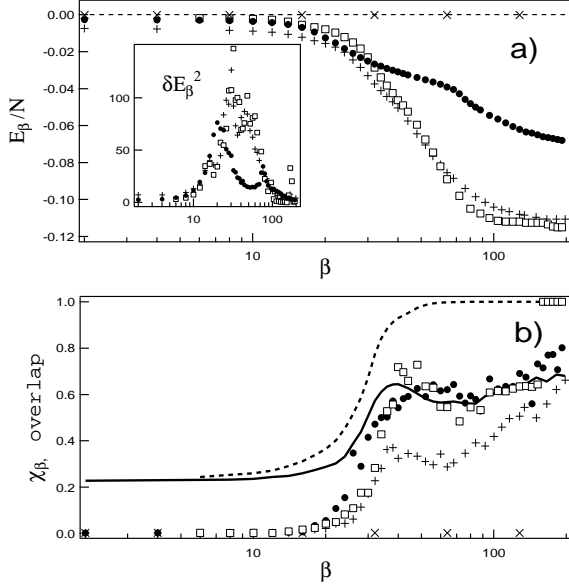


FIG. 2. **a)** Energy  $E_\beta$  as a function of  $\beta$  for random ( $\times$ ), S&P500 (+) and correlated ( $\square$ ) data sets of length  $D = 1599$  respectively. The results for the S&P500 data set over the last  $D = 400$  days are also shown ( $\bullet$ ). Inset: square energy fluctuation  $\delta E_\beta^2$  vs  $\beta$  for the same data sets (same symbols). **b)** Autocorrelation  $\chi_\beta$  as a function of  $\beta$  for the same data sets (same symbols). The full (dashed) line refers to the overlap with the configuration  $s^*$  for the S&P500 (correlated) data set with  $D = 1599$ .

We monitored these quantities for three different data sets all composed by  $N = 443$  time series: 1) uncorrelated time series; 2) the time series of daily returns of the assets composing the S&P500 index [7,12] 3) correlated time series generated by Eq. (2) with given  $s_i = s_i^*$  and  $g_s = g_s^*$ . The first and the third data sets serve to test the method in cases where we know the answer.

Let us start with a truly uncorrelated time series with  $D = 1599$ . We compute  $C_{i,j}$  and study the corresponding Hamiltonian  $H_c$  by the MC method. We do not expect any clustering to emerge in this case. Indeed, the internal energy  $E_\beta$  stays very close to 0 (see Fig. 2a) for all values of  $\beta$  investigated up to  $\beta = 512$ . Correspondingly no persistent cluster arises, i.e.  $\chi_\beta \simeq 0$ .

The results change turning to correlated data. Let us first discuss the S&P500 data for  $D = 1599$ : As Fig. 2a shows, for  $\beta \approx 20$  the energy  $E_\beta$  starts deviating significantly from zero. For  $\beta > 20$  persistent clusters are present:  $\chi_\beta$  rapidly raises from zero and it has a maximum at  $\beta \approx 40$  (see Fig. 2b). The energy fluctuations reported in the inset shows a broad peak of intensity marking the onset of an ordered low temperature phase. As  $\beta$  increases the dynamics is significantly slowed down. At  $\beta \approx 200$  the energy reaches a minimal value

$E_\beta \simeq -0.11N$  and does not decrease significantly increasing  $\beta$  at least up to  $\beta = 4095$ . This energy is smaller than that of the ferromagnetic state ( $E_f = -0.086N$ ), with all sets in the same cluster. The system in this range of temperatures visits only few configurations.

The statistical properties of cluster configurations, as  $\beta$  varies, are shown in Fig. 3. For small  $\beta$  only small clusters survive to thermal fluctuations. As  $\beta$  increases a distribution of cluster sizes develops. At low temperatures the rank order plot of  $n_s$  reveals a broad distribution of clusters with the largest aggregating more than 190 sets. By a power law fit of this distribution, we find that the number of clusters with more than  $n$  sets decays as  $n^{-0.83}$ . The scatter plot of  $c_s$  versus  $n_s$  also reveals a non-trivial power law dependence  $c_s \sim n_s^{1.66}$ . This gives a statistical characterization of the dominant configurations of clusters at low energy [13].

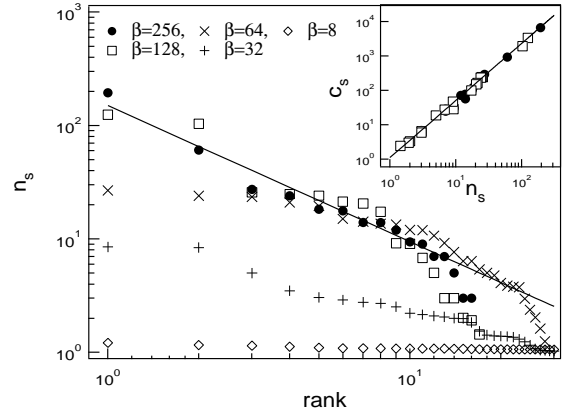


FIG. 3. Rank plot of  $n_s$  for several values of  $\beta$ . The line corresponds to  $n \sim \text{rank}^{-1.2}$ . Inset:  $c_s$  versus  $n_s$  for  $\beta = 256$  ( $\bullet$ ) and  $\beta = 128$  ( $\square$ ). The line corresponds to  $c/sim n^{1.66}$ .

For  $D = 400$  we find two transitions at  $\beta_1 \approx 20$  and at  $\beta_2 \approx 80$  which are signalled by bending in the  $E_\beta$  curve and by peaks in the  $\delta E_\beta^2$  vs  $\beta$  plot. At the first temperature clusters start to appear. For  $\beta < \beta_2$  the largest cluster groups less than 30 sets and for  $\beta > \beta_2$  larger clusters  $n_s \approx 100$  appear. This hints at a time dependence of correlations, which are averaged in the  $D = 1599$  data set. For even shorter time series we found that sampling errors, acting like a temperature, destroy large clusters and only relatively small clusters ( $n_s < 40$  for  $D = 60$ ) were found.

We build a syntetic correlated data set of  $D = 1599$  points using Eq. (2) with a structure  $\mathcal{S} = \mathcal{S}^*$  and parameters  $\mathcal{G} = \mathcal{G}^*$ . The structure  $\mathcal{S}^*$  is a typical low energy configuration for the S&P500 data set for  $D = 1599$ . The parameters  $g_s^*$  were deduced from the  $n_s$  and  $c_s$  of this configuration, via Eq. (4). The distribution of eigenvalues of  $C_{i,j}$  is shown in Fig. 1a. This data set is useful for at least two reasons: first it allows one to understand to what extent a struture of correlation put by hand with the form dictated by Eq. (2) can be correctly recovered.

Secondly it allows us to compare the results found for the S&P500 data with those of a time series with correlations described by Eq. (2), of a similar nature.

For  $\beta < 150$ , the behaviors of  $E_\beta$ ,  $\delta E_\beta^2$  and  $\chi_\beta$  are similar to those found for the S&P500 data (see Fig. 2). A second, sharp peak in  $\delta E_\beta^2$  at  $\beta \approx 170$  signals a new clustering phase transition. Below this temperature, as shown by the plot of  $\chi_\beta$  (fig. 2b), the MC dynamics freezes into the original structure  $\mathcal{S}^*$ . The overlap with the configuration  $\mathcal{S}^*$ , defined as in Eq. (6) as the fraction of “bonds”  $s_i = s_j$  for which  $s_i^* = s_j^*$ , quickly converges to 1 (see fig. 2b) for the synthetic time series, whereas it remains around 60% for the S&P500 data set. This, on one hand means that the original structure  $\mathcal{S}^*$  can be recovered quite efficiently. On the other hand, it suggests that several cluster configuration compete at low temperatures in the S&P500 data set.

Eq. (2) with a single cluster configuration ( $\beta \rightarrow \infty$ ), is inadequate to capture the full complexity of the correlations in the S&P500 data set. Probabilistic clustering, where several cluster structures  $\mathcal{S}$  are allowed with their Gibbs probability  $P\{\mathcal{S}\}$  (and finite  $\beta$ ) provides a much better approximation. At finite  $\beta$  each set  $i$  may belong to several clusters and we can measure the corresponding coupling strenght  $g_{s,i}(\beta) = \langle \hat{g}_s \delta_{s,s_i} \rangle_\beta$ . Taking these as the parameters of the generalized model

$$\xi_i(d) = \frac{\sum_s \sqrt{g_{s,i}(\beta)} \eta_s(d) + \epsilon_i(d)}{\sqrt{1 + \sum_s g_{s,i}(\beta)}}, \quad (7)$$

we can build an artificial time series  $\vec{\xi}_i$  and compute the correlation matrix  $C_{i,j}^{(\beta)}(D)$ . Here  $\beta$  is a free parameter which can be adjusted to “fit” the S&P500 correlation matrix. The eigenvalue spectra of the two matrices are compared in fig. 1b for  $\beta = 48$ . The value of  $\beta$  was chosen by visual inspection as that giving the best fit. The curves are remarkably close, suggesting that Eq. (7) provides a good statistical description of the correlations among assets. Fig. 1b also shows the *noise undressed* matrix  $C_{i,j}^{(\beta)}(\infty)$ , which allows one to appreciate the effect of noise dressing. As expected, noise mainly affects small eigenvalues.

The applicability of the method can be extended considerably to a generic data set  $\{\vec{x}_i\}_{i=1}^N$ .  $\vec{x}_i$  need not be a time series. The distribution of  $x_i(d)$  need not be Gaussian and it does not even need to be the same across  $i$ . For example,  $x_i(d)$  may be the measure of the  $d^{\text{th}}$  feature of the  $i^{\text{th}}$  object or the concentration of species  $i$  in the  $d^{\text{th}}$  sample of an experiment. The idea is to map the data set  $\vec{x}_i$  into a Gaussian time series  $\vec{\xi}_i$  to which we apply Eq. (2). The mapping results from requiring that non-parametric cross-correlations  $\tau_{i,j}^x = \tau_{i,j}^\xi$  are preserved. To do this in practice we compute Kendall’s  $\tau$  [14] for the  $\vec{x}_i$  data sets:  $\tau_{i,j}^x = \langle \text{sign}[x_i(d) - x_i(d')] \text{sign}[x_j(d) - x_j(d')] \rangle_{d < d'}$ . We note that, for two infinite Gaussian time series with cor-

relation  $c$  we have  $\tau = \frac{2}{\pi} [\tan^{-1} \sqrt{\frac{1+c}{1-c}} - \tan^{-1} \sqrt{\frac{1-c}{1+c}}]$ . Inverting this relation, we find the correlation  $c = C_{i,j}$  as a function of  $\tau = \tau_{i,j}^\xi = \tau_{i,j}^x$ . This allows us to build the Hamiltonian which can then be studied.

With respect to ref. [2], our approach does not need any assumption on the form of the Hamiltonian. As input, the method only needs the correlation matrix  $C_{i,j}$  (or  $\tau_{i,j}$ ). The range of interactions is set by the correlations themselves. For small  $D$ , the local interaction of ref. [2] may well be more efficient in capturing the structure of data. Our method is most useful in cases where  $D \sim N \gg 1$ . These ideas can clearly be extended to models of correlations different from Eq. (2).

I acknowledge R. Zecchina, R. Pastor-Satorras and L. Giada for interesting discussions and R. N. Mantegna for providing the S&P500 data.

- 
- [1] K. Rose, E. Gurewitz and G. C. Fox, Phys. Rev. Lett. **65**, 945 (1990).
  - [2] M. Blatt, S. Wiseman and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996); S. Wiseman, M. Blatt and E. Domany, Phys. Rev. E **57**, 3767 (1997).
  - [3] L. Laloux *et al.* Phys. Rev. Lett. **83**, 1467 (1999); V. Plerou *et al.*, Phys. Rev. Lett. **83**, 1471 (1999).
  - [4] R. N. Mantegna, Eur. Phys. J. B **11**, 193 (1999).
  - [5] L. Kullmann, J. Kertesz, R. N. Mantegna, e-print cond-mat/0002238.
  - [6] E.J. Elton, M.J. Gruber, *Modern Portfolio Theory and Investment Analysis* (J.Wiley & Sons, New York, 1995).
  - [7] The data covers the period July 3<sup>rd</sup> 1989 to December 27<sup>th</sup> 1995).  $\xi_i^n(d) = \frac{\log p_i(d)/p_i(d-1) - r_i}{\sigma_i}$  where  $p_i(d)$  is the price of asset  $i$  in day  $d$ . The parameters  $r_i$  and  $\sigma_i$  are fixed so that  $\sum_d \xi_i(d) = 0$  and  $\sum_d \xi_i^2(d)/D = 1$  for all  $i$ .
  - [8] A. M. Sengupta, P. P. Mitra, Phys. Rev. E **60**, 3389 (1999).
  - [9] J. D. Noh, e-print cond-mat/9912076.
  - [10] In principle, the low temperature properties may be dominated by exponentially many meta-stable states, as in spin glasses. This would mean that the *ansatz* (2) is not appropriate. Indeed multiple solutions is an undesirable outcome of a statistical fit. In the cases discussed here, we found no glassy behavior.
  - [11] N. Metropolis *et al.*, J. Chem. Phys. **21**, 1087 (1953). Note that cluster algorithms, as those used in ref. [2], cannot be used for  $H_c$ .
  - [12]  $\xi_i(d)$  is not Gaussian for the S&P500 data set [see R. N. Mantegna, H. E. Stanley, Nature **376**, 46 (1995)]. However this affects very weakly the matrix  $C_{i,j}$  for  $D \gg 1$ .
  - [13] From the conomic point of view, we find reasonable results: companies in the same sector belong to the same cluster. We shall not discuss these issue here, but rather focus on the clustering method.
  - [14] See e.g. W. H. Press *et al.* *Numerical recipes in Fortran*, p. 633 (Cambridge University Press, 1992).